Research topic: Influence of YouTube video characteristics and creator demographics on viewer engagement and sales performance

General description:

The dataset I am working with is from Kaggle. The dataset consists of information about YouTube videos and the creators of those videos respectively. It contains information on 21 variables for 1096 videos. It was collected in 2022 by a bunch of business analyst interns from a company called KultureHire to figure out the success ratio of YouTube content creators. Each row of the dataset corresponds to a different YouTube video, and each column reports on the value of a certain variable for a given video. These variables include YouTube video characteristics (duration of the video, language, hashtags), creators demographics (total subscribers, gender), viewer engagement metrics (number of likes, comments, etc). There is no known limitations or issues with the dataset.

Caution:

I have removed potential outliers from this dataset before proceeding with the analysis. The outliers are a result of the linear regression model between the predictor 'Total channel subscribers' and the outcome 'No of likes'. The outliers are calculated by the 2RMSE method, which removes the observations that are more than 2 mean square errors away from the regression line. There are 1069 observations after removing outliers.

My main research question:

Which YouTube video and creator characteristics are the strongest predictors of high viewer engagement, suggesting potential sales success?

Understanding the characteristics of a YouTube videos and the creators of those videos that predict high viewer engagement is crucial for marketers and content creators who aim to maximize the impact of their digital marketing strategies. Higher engagement often is directly proportional to increased consumer interest and potential sales success. The findings of this research can also help creators tailor their content to better meet audience preferences.

Hypothesis about the main relationships of interest.

Total Channel Subscribers:

Null Hypothesis: There is no relationship between a YouTube channel's total subscribers and the number of likes the video in that channel receives.
Alternative Hypothesis: There is a relationship between a YouTube channel's total subscribers and the number of likes the video in that channel receives.

Duration of the video:

Null Hypothesis: There is no relationship between the duration of a YouTube video and the number of likes the video receives.
Alternative Hypothesis: There is a relationship between the duration of a YouTube video and the number of likes the video receives.

Subtitles:

Null Hypothesis: There is no difference in the number of likes received between YouTube videos with and without subtitles.
Alternative Hypothesis: There is a difference in the number of likes received between YouTube videos with and without subtitles.

Maximum quality of the video:

Null Hypothesis: There is no relationship between the maximum quality of a YouTube video and the number of likes the video receives.
Alternative Hypothesis: There is a relationship between the maximum quality of a YouTube video and the number of likes the video receives.

Premiered or not:

Null Hypothesis: There is no difference in the number of likes received between premiered YouTube videos and non-premiered YouTube videos.

Alternative Hypothesis: There is a difference in the number of likes received between premiered YouTube videos and non-premiered YouTube videos.

Interpretation:

- Rejecting the null hypothesis in any of these tests suggests that there is evidence to support the alternative hypothesis. For example, rejecting the null hypothesis for total channel subscribers would imply that as the number of subscribers increase, the viewer engagement (no of likes) can either increase or decrease. Thereby providing insights what factors influence the viewer engagement, and potentially improving sales success.
- Statistical significance indicate that the observed relationship is unlikely due to a random chance. A significant finding from these tests suggests that changes in these predictors are associated with changes in the number of likes a video can receive.
- A lack of statistical significance (failure to reject null hypothesis) can be a cause for concern as it could mean that the sample data is not strong enough to decide whether there's a relationship between the predictors and the outcome variable. There could be other factors that influence the outcome.

Descriptive statistics of the important variables:

Viewer engagement metrics:

	Mean	*Median*	*S.D.*	*Min*	*Max*
NoofLikes	*2.782e+05*	*36000*	*7.033e+05*	*0.000*	*6.400e+06*
NoofComments	*27961*	*1400*	*3.234e+05*	*0.000*	*7.380e+06*

Creator and video demographics:

	Mean	*Median*	*S.D.*	*Min*	*Max*
TotalChannelSubc~	*1.218e+07*	*2.500e+06*	*2.229e+07*	*34.00*	*2.250e+08*
DurationofVideo	*0.06931*	*0.008194*	*0.1721*	*0.000*	*2.160*
MaximumQualityof~	*1252*	*1080*	*431.6*	*240.0*	*2160*

Proportions of the important categorical variables (total observations – 1069):

Total channel subscribers (>1,000,000; <= 1,000,000):

- YouTube channels that have more than a million subscribers - 64.08%
- YouTube channels that have less than a million - 35.92%

Subtitle:

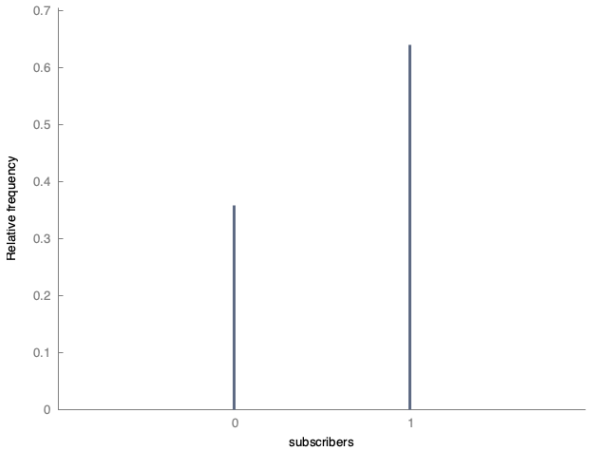
- With subtitles – 60.90%
- Without subtitles – 39.10%

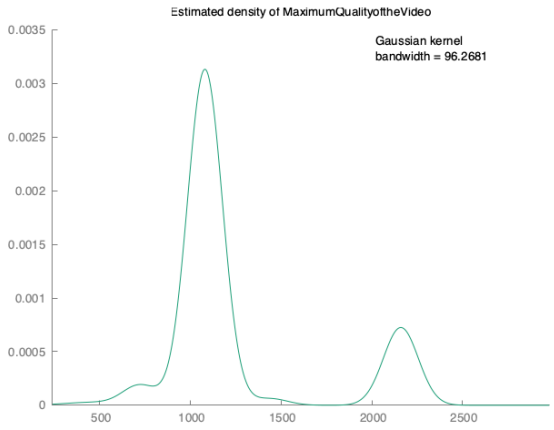
Maximum quality of the video:

- 240px - 0.09%
- 360px - 0.28%
- 480px - 0.47%
- 720px - 4.58%
- 1080px - 75.58%
- 1440px - 1.50%
- 2160px - 17.49%

From the descriptive statistics, we can see that the data consists of around 64.08% of videos posted by creators who have more than a million followers and around 35.92% posted by creators who have less than a million followers. The maximum quality of the video ranges from 240px to 2160px where a majority of those have a maximum quality of 1080px. We are interested in studying any relationships that exist between the viewer engagements and the video characteristics, and the creator demographics variables.

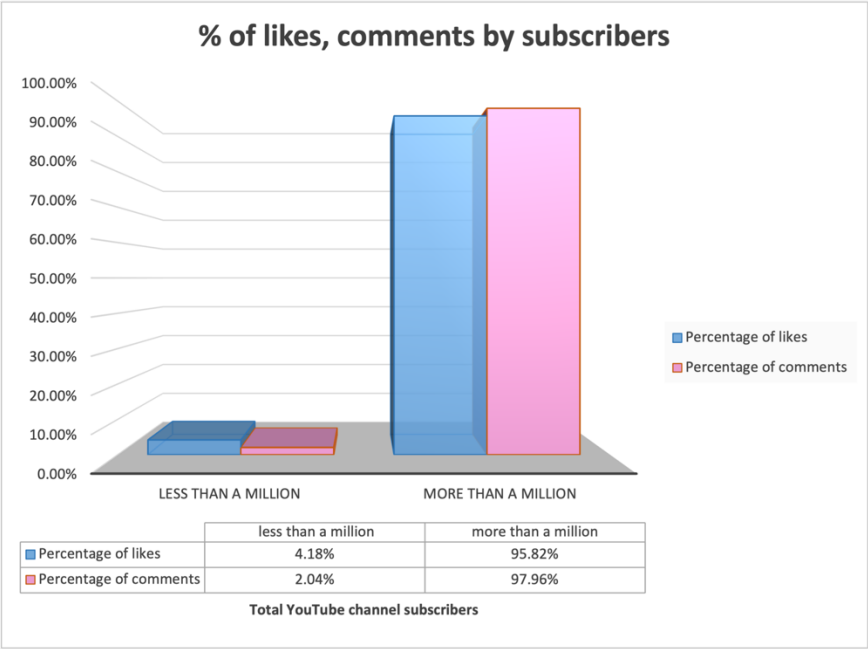
Graphs and density plots:





The frequency distribution of the variable 'subscribers' where 0 denotes '< 1000000' subscribers and 1 denotes '>1000000' subscribers. From this sample, we can see that the proportion of videos whose creators have more than a million subscribers are higher than the creators who have less than a million subscribers. The estimated density plot of the variable 'MaximumQualityoftheVideo' shows that most number of YouTube videos have a maximum quality of 1080px, as there is a high peak in that area.

% of likes and comments by total channel subscribers:



The YouTube videos posted by creators who have more than a million subscribers tend to be performing well in terms of likes and comments when compared to the videos posted by creators who have less than a million subscribers.

Regression models:

Simple linear regression model with the main predictor variable:

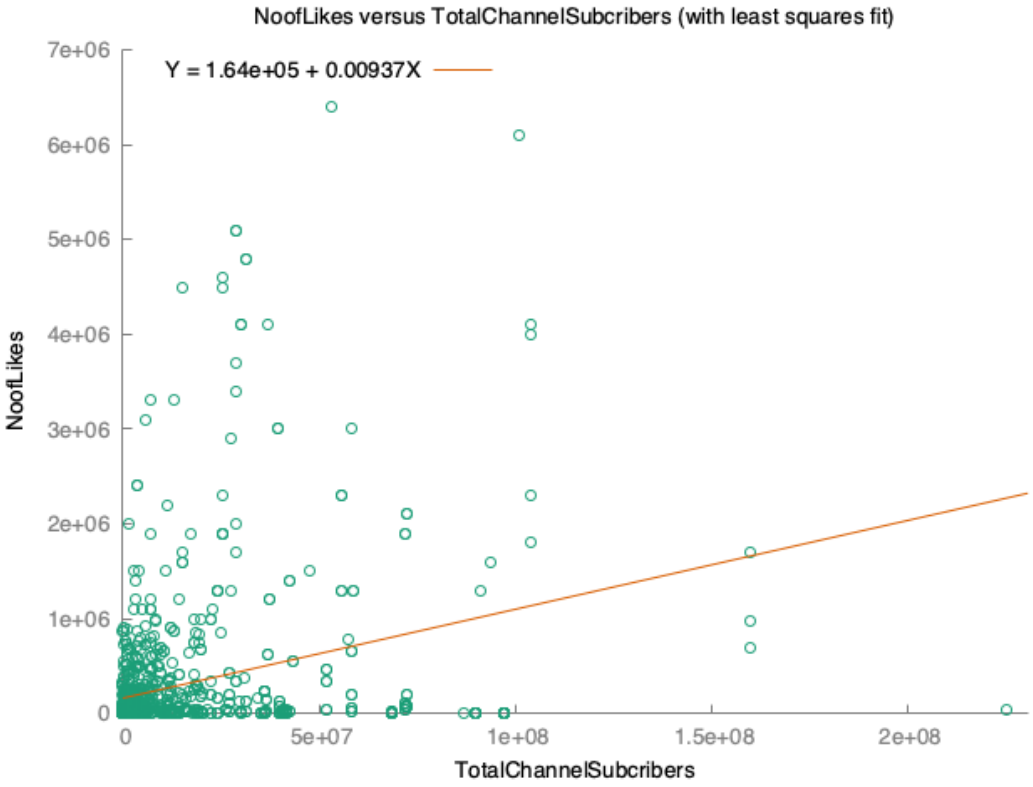
NoofLikes yhat = 164118 + 0.00937(TotalChannelSubcribers)

R-squared = 0.088105

Adjusted R-squared = 0.087250

	coefficient	*std. error*	*t-ratio*	*p-value*
const	*164118*	*23421.1*	*7.007*	*4.30e-12*

TotalChannelSubc~	*0.00936709*	*0.000922561*	*10.15*	*3.48e-23*



Interpretations:

- The slope coefficient value 0.00937 indicates that for each additional subscriber, the number of likes can increase by 0.0094.
- The R squared value 0.088 shows that only about 8.8% of the total variation in the number of likes can be explained by the total channel subscribers.
- The high t-statistics and extremely low p-values indicate a positive linear relationship between the variables ‘no of likes’ and ‘total channel subscribers’.

Importance and possible policy implications:

Though the low R-squared and adjusted R-squared values suggest that only about 8.8% of the variation in the viewer engagement is explained by the number of subscribers, it still is a positive coefficient. This means that content creators can focus on improving their YouTube channel’s subscribers which will lead to higher viewer engagement.

For platforms and policymakers, this means that supporting tools and strategies that help content creators attract and retain subscribers.

Multiple regression model including all the important predictors:

NoofLikes yhat= 136128 + 0.00981(TotalChannelSubcribers) + 502355(DurationofVideo) + 9643.20(Subtitle) - 78.1(MaximumQualityoftheVideo) + 65471.6(PremieredorNot)

R-squared = 0.108402

Adjusted R-squared = 0.104209

	coefficient	*std. error*	*t-ratio*	*p-value*
const	*136128*	*113161*	*1.203*	*0.2293*
TotalChannelSubc~	*0.00980873*	*0.000924430*	*10.61*	*4.54e-25*
DurationofVideo	*502355*	*121333*	*4.140*	*3.74e-05*
Subtitle	*9643.20*	*42151.0*	*0.2288*	*0.8191*
MaximumQualityof~	*−78.0765*	*47.4938*	*−1.644*	*0.1005*
PremieredorNot	*65471.6*	*67168.5*	*0.9747*	*0.3299*

Interpretations:

B1 – The slope coefficient value 0.0098 indicates that for each additional subscriber, the number of likes increase by 0.0098, holding other factors constant.

B2 - The slope coefficient value 502355 indicates that for every additional second of the video, the number of likes increase by approximately 502,355, holding other factors constant.

B3 - The slope coefficient value 9643.20 indicates that videos with subtitles receive 9,643 more likes compared to the videos without subtitles, holding other factors constant.

B4 – The slope coefficient value -78.1indicates that for each additional increase in the maximum quality of the video resolution, the number of likes decrease by approximately 78, holding other factors constant.

B5 – The slope coefficient value 65471.6 indicates that videos that were premiered receive 65,472 more likes compared to the videos that weren't premiered, holding other factors constant.

The model's R squared value is 0.1084. Approximately, 11% of the total variation in the number of likes is explained by the combined effect of these predictors included in this model.

Statistical significance of the estimated coefficients:

- The variables 'TotalChannelSubcribers' and 'DurationofVideo' have p-values <0.05, indicating a significant relationship with 'no of likes'.
- The variables 'subtitle', 'MaximumQualityoftheVideo', and 'PremieredorNot' have p-values >0.05, indicating no significant relationship with 'no of likes'.

Comparing of adjusted R-squared values between the models:

This model's adjusted R-squared value is 0.104. Approximately, 10.4% of the total variation of a YouTube video likes is explained by the combined effect of the variables, after adjusting for the number of predictors.

The previous model's adjusted R-squared value is 0.0873. Comparing the adjusted R-squared values shows an increase when including these predictors , indicating they add some explanatory power.

F-test to test the overall significance of this model:

$$(5, 1063) = 25.84838; \quad P\text{-value}(F) = 1.11e\text{-}24$$

Null hypothesis Ho : b1 = b2 = b3 = b4 = b5 =0
Alternative hypothesis Ha: at least one beta does not = 0

The p-value **1.11e-24 < 0.05** → Reject the null hypothesis. This indicates that the overall model is statistically significant even though the addition of the variables 'subtitle', 'MaximumQualityoftheVideo' and 'PremieredorNot' are not significant. This model can prove to be of value in predicting the factors influencing the YouTube viewer engagement and potential sales success.

Importance and possible policy implications:

Increase in the number of a YouTube channel's subscribers, longer videos, inclusion of subtitles and premiering videos are contributing factors that boost viewer engagement which then leads to potential

sales success. Advertisers can leverage insights on the length of the video, and premiering strategies to target content that maximizes the engagement.

The negative effect for the coefficient of the variable 'MaximumQualityoftheVideo' indicate that it is not necessary for a video to be in the maximum quality possible to improve viewer engagements. Let us explore more on this by performing a multinomial logit model analysis,

Multinomial logit model (MaximumQualityoftheVideo, NoofLikes):

	coefficient	*std. error*	*z*	*p-value*	
MaximumQualityoftheVideo = 360					
const	*1.40890*	*1.25949*	*1.119*	*0.2633*	
NoofLikes	*−3.81924e-06*	*6.98090e-06*	*−0.5471*	*0.5843*	
MaximumQualityoftheVideo = 480					
const	*1.62325*	*1.17980*	*1.376*	*0.1689*	
NoofLikes	*−7.35668e-08*	*2.27036e-06*	*−0.03240*	*0.9742*	
MaximumQualityoftheVideo = 720					
const	*3.80585*	*1.08778*	*3.499*	*0.0005*	***
NoofLikes	*3.38276e-07*	*2.04463e-06*	*0.1654*	*0.8686*	
MaximumQualityoftheVideo = 1080					
const	*6.62904*	*1.07744*	*6.153*	*7.62e-10*	***
NoofLikes	*2.73555e-07*	*2.03778e-06*	*0.1342*	*0.8932*	
MaximumQualityoftheVideo = 1440					
const	*2.79277*	*1.11004*	*2.516*	*0.0119*	**
NoofLikes	*−1.09694e-07*	*2.11970e-06*	*−0.05175*	*0.9587*	
MaximumQualityoftheVideo = 2160					
const	*5.22923*	*1.07964*	*4.844*	*1.28e-06*	***
NoofLikes	*9.51205e-09*	*2.04242e-06*	*0.004657*	*0.9963*	

The coefficients of the max quality levels 720px, 1080px and 2160px exhibit a positive effect on the number of likes while the other quality levels exhibit a negative effect. This model indicates that YouTube videos with a maximum quality more than 720px are expected to have higher log-odds of increasing the number of likes.

Alternative nonlinear functional models:

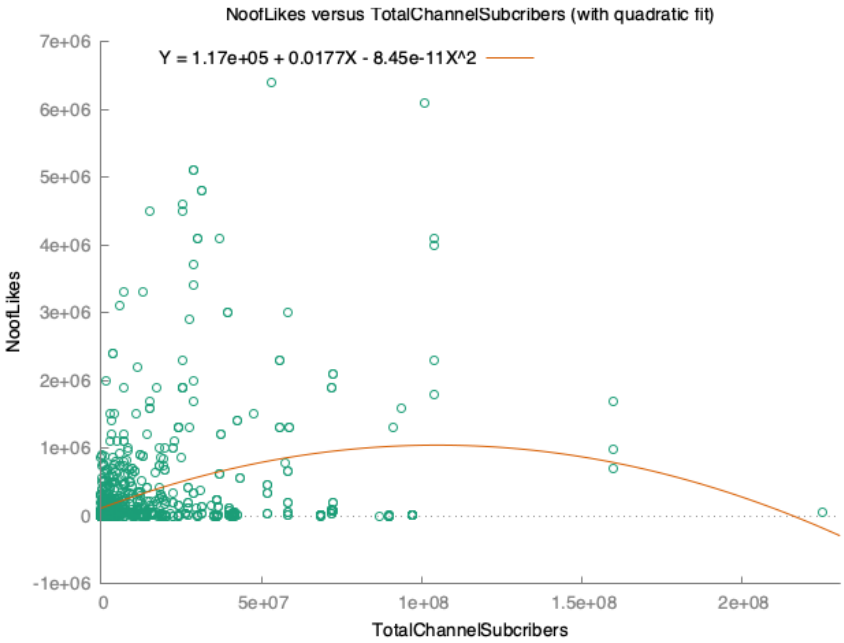
Quadratic model:

NoofLikes yhat = 116726 + 0.0177(TotalChannelSubcribers) - 8.45e-11(TotalChannelSubcribers^2)

R-squared = 0.112200

Adjusted R-squared = 0.110534

	coefficient	*std. error*	*t-ratio*	*p-value*
const	*116726*	*24742.5*	*4.718*	*2.70e-06*
TotalChannelSubc~	*0.0177320*	*0.00180220*	*9.839*	*6.35e-22*
Subs2	*−8.45160e-11*	*1.57129e-11*	*−5.379*	*9.21e-08*



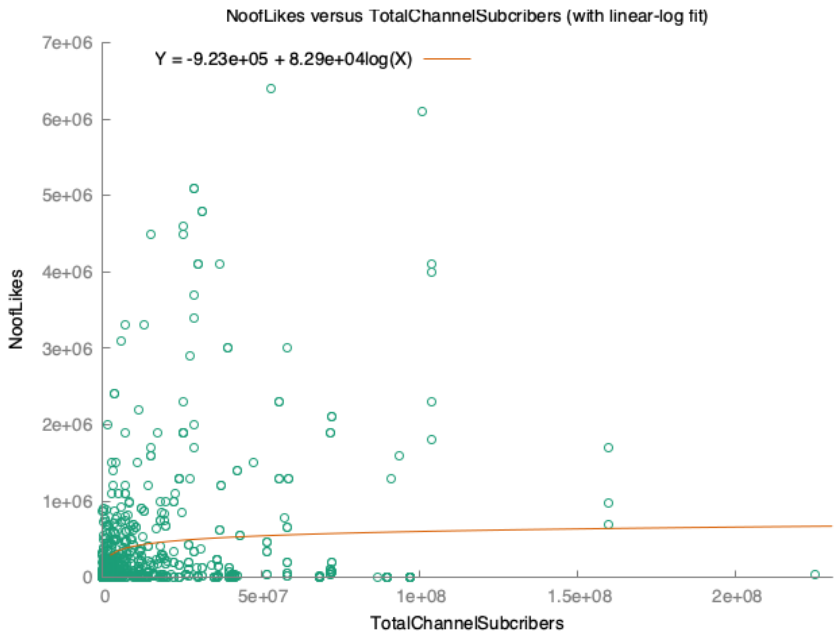
Linear-log model:

NoofLikes yhat = −922557 + 82929.7(log(TotalChannelSubscribers)

R-squared = 0.089500

Adjusted R-squared = 0.088647

	coefficient	*std. error*	*t-ratio*	*p-value*
:onst	*−922557*	*119030*	*−7.751*	*2.12e-14*
nsub	*82929.7*	*8097.59*	*10.24*	*1.53e-23*



Comparison between the quadratic and the linear log model:

Though the R-squared and the adjusted R-squared values are higher in the quadratic model compared to the linear log model, indicating strong statistical fit. There is a diminishing return, where the number of likes decreases as the number of total subscribers increase in the quadratic model. The linear log model seems to be an appropriate fit among these two, as there is a consistency in the rate of number of likes without any diminishing returns.

Let us also explore an interaction model as the above models do not prove to strong fits.

***Interaction model**:*

*NoofLikes yhat = 147116 + 0.00678(TotalChannelSubcribers) + 0.0987(TotalChannelSubscribers * DurationoftheVideo)*

R-squared = 0.210272

Adjusted R-squared = 0.208790

	coefficient	*std. error*	*t-ratio*	*p-value*
const	*147116*	*21846.3*	*6.734*	*2.69e-11*
TotalChannelSubc~	*0.00677608*	*0.000882324*	*7.680*	*3.59e-14*
subs_duration	*0.0986546*	*0.00768246*	*12.84*	*3.36e-35*

Interpretations:

- B1 – The slope coefficient value 0.0068 indicates that for each additional subscriber, the number of likes increase by 0.0068, holding other factors constant.
- B2 - The slope coefficient value 0. 099 indicates that for each additional subscriber, and a additional second of a video simultaneously, the number of likes increase by 0.099, holding other factors constant.

With an adjusted R-squared value of 0.20, this interaction model proves to be a better fit statistically compared to the previous nonlinear models. In theory, this means that the viewer engagement improves when a YouTube channel's total subscribers and the length of the video increase, simultaneously.

Importance and policy implications:

From this model, we can understand how the number of a YouTube channel's subscribers and a video's duration jointly influence the viewer engagement. Also, about 20% of the total variation in the number of likes is explained by this combined effect of predictors.
Content creators can consider optimizing the video length but also should make sure that it is aligned with their subscriber preferences to maximize their engagement rates.

Important takeaways from this research:

- After testing several models, the interaction model shows that the combined effect of **total subscribers and video duration** explains about 20% of the variation in the number of likes a YouTube video can receive. This suggests that optimizing video length in alignment with subscriber preferences can enhance engagement.
- Content creators aiming to maximize viewer engagement can consider uploading their videos with a **resolution of 720px** or more, **preferred quality is 1080px**. Videos with better resolution offer a superior viewing experience which can lead to higher number of likes.
- **Adding subtitles and premiering a YouTube video** can increase viewer engagement. Videos with subtitles and premieres tend to receive more likes than those without.

- **YouTube videos with over a million channel subscribers** tends to attract more likes and comments. So, content creators can focus on techniques to grow their subscribers count by posting valuable content, marketing, and advertising, etc.

Limitations:

One of the creator demographics 'CreatorGender' seems to be inconsistent in this dataset with a lot of values with 'N/A' and blanks. This is a shortcoming so I have not used this in my analysis.

Scope for further analysis:

- To analyze the remaining 80% of the variation in viewer engagement (number of likes or comments) not explained by current models, exploring additional potential predictors like total channel views, video description, date of upload, number of videos in a channel can be helpful.
- Collect additional data if possible, generate and test hypotheses and continuously refine models based on new insights.